Communications in Statistics—Simulation and Computation®, 43: 1595–1610, 2014

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.738840

Taylor & Francis
Taylor & Francis Group

Variable-Selection-Based Epidemic Disease Diagnosis

XINXING ZHOU, KAIBO WANG, AND LEI ZHAO

Department of Industrial Engineering, Tsinghua University, Beijing, China

Epidemic surveillance in a community involves monitoring infection trend, triggering alarms before outbreaks, and identifying sources and paths of disease transmission. Algorithms for outbreak detection that are derived from industrial statistical process control (SPC) and scan statistics have been reported in the literature, but there are relatively few methods reported for identifying transmission paths. In this work, we propose an expanded spatial-temporal (EST) model for identifying infection sources. Three dimensional information, subject, location, and time, are expanded into a two-dimensional space by dividing the time horizon into segments and multiplying each segment by the locations. Based on the EST model, we further propose a variable-selection algorithm to identify potential location/time combinations as sources of infection, and thus achieve diagnosis. Numerical simulations show that the proposed scheme is effective in locating infection sources.

Keywords Diagnosis; Healthcare surveillance; Statistical process control; Variable selection.

Mathematics Subject Classification Primary 62H20; 62J05 Secondary 62P30.

1. Introduction

Globalization has heightened international trade and travel, which in turn have increased both the scale and speed of epidemic disease dissemination (Cash and Narasimhan, 2000). Recent epidemic diseases, such as SARS (Yip et al., 2008) and bird flu, disseminated globally within a short period of time. The most recent instance, the pandemic H1N1 2009, attacked more than 214 countries, overseas territories and communities worldwide, leading to over 18,036 deaths prior to May 9, 2010 (WHO, 2010). It is commonly realized that efficient epidemic surveillance can help control the spread of infectious diseases, save lives, and reduce losses (see, e.g., Allard, 1998; Cash and Narasimhan, 2000; Woodall et al., 2008).

In recent years, great social and academic efforts have been devoted to epidemic surveillance activities. There are two specific tasks that are of great importance: outbreak detection and source diagnosis. By closely monitoring epidemic data, outbreak detection aims to trigger alarms before a real and serious outbreak occurs; on the other hand, source diagnosis aims to identify sources of disease infection, locate high-risk locations and provide information for effective control measures. Therefore, extensive studies aimed at avoiding

Received November 17, 2011; Accepted October 8, 2012
Address correspondence to Dr. Kaibo Wang, Department of Industrial Engineering, Tsinghua University, Beijing 100084, China; E-mail: kbwang@tsinghua.edu.cn

serious outbreaks and controlling disease dissemination have been conducted in collecting and analyzing epidemic-related data, especially for the purpose of outbreak detection (see, e.g., Allard, 1998; Andersson, 2009; Rolka et al., 2007; Tsui et al., 2008).

Different methods, such as statistical process control (SPC) and scan statistics, have been developed for detecting outbreaks in complex healthcare data streams. SPC, or more specifically, control chart techniques, have been widely used in the manufacturing industry to monitor the process status and detect faults. Because the detection of epidemic outbreaks is similar to the detection of process faults, traditional SPC techniques, including exponentially weighed moving average (EWMA), cumulative sum (CUSUM), and Hotelling's T^2 charts, have been extended for use in healthcare surveillance (see, e.g., Allard, 1998; Carey, 2003; Frisen and Wessman, 1999; Hanslik et al., 2001). These methods are generally effective in detecting sustained shifts in a process and are potentially helpful in detecting epidemic outbreaks. However, many researches have pointed out that epidemic data, different from the data collected in industrial processes, often consist of both spatial (incidence location) and temporal (incidence time) information (Hutwagner et al., 2003; Pfeiffer et al., 2007). Most conventional SPC methods are designed based on temporal information only. In addition, most of these SPC methods cannot provide diagnostic information to help identify infection sources and outbreak locations when an alarm is triggered.

The typical method for identifying outbreak sources involves the tracking of laboratory-confirmed cases. When the primary case is determined based on the manifestation time of symptoms, all related subjects in the social network of the case are investigated. After collecting the demographic information and activities of related subjects, the hypothesis is tested to evaluate the significance of potential sources. For example, to identify the source of a food-poisoning outbreak, Cowden et al. 1989, collected data on the food consumed one week before the onset of the symptoms. The numbers of subjects who have or have not consumed the food are separately counted. Then, statistical tests are conducted to confirm whether a statistically significant number of patients are associated with this food. Horby et al. 2003, reported the use of a similar approach in analyzing a national outbreak of multi-resistant Salmonella enterica. However, this method of dealing with such detailed data is very inefficient and may ignore important signals when the data are very sparse.

Recently, algorithms based on scan statistics that can effectively use both temporal and spatial information have been developed for the surveillance purpose (Kulldorff, 2001; Sonesson and Bock, 2003). The scan statistics consider both temporal and spatial information, and involve counting the number of incidences in different regions during the most recent time period with a fixed length (Woodall et al., 2008). Scan-based methods can detect changes in infectious rates at an early stage and provide diagnostic information relating to both time and location (Shmueli and Burkom, 2010). However, scan statistics only use spatial information at certain aggregated levels; the total number of infected subjects in each area is counted and monitored. More detailed information of the subjects, such as the travelling path and time information, is not considered.

In epidemic surveillance, diagnosis involves finding the sources, time and location of infection, and determining the paths of transmission. Epidemic disease transmission usually begins in clustered-population communities (Bailey, 1986). Providing accurate diagnosis information is as important as triggering alarms before outbreaks. The capability to identify high-risk locations where a subject may be infected and to determine infection locations if one is already infected is critical to an epidemic surveillance system. Incorrectly identified times and locations may result in ineffective immunity efforts, or may even void such efforts. Hence, diagnosis is critical to healthcare surveillance and control. Traditional SPC methods may not be able to provide adequate diagnosis information because these methods

only use temporal information. As was suggested by Tsui et al. 2008, the scan-based methods can provide certain positional information when an alarm is triggered, but the diagnostic information remains at an aggregated level, without details of the subjects and their activities.

Detailed activities and corresponding location and time information of a subject are known to contain rich information that may be related to disease transmission. Such information can be gained under some special situations. Hu et al. 2004, reported the efforts they made in the 2003 SARS outbreak. The authors tracked the recent activities of confirmed cases, and isolated the people and locations they visited. In this way, detailed activities and corresponding location and time information are collected. In a population-intensive campus, the attendance records, dinning and consuming records, and curriculum or task allocation information collected automatically by an IT system could be used for such diagnosis purpose.

Recently, Horby et al. 2003, suggested that the use of such information for surveillance can effectively aid in accurate source diagnosis. Therefore, in this paper, we propose a new diagnosis method for healthcare surveillance by considering the trajectories of an individual subject in a particular area. We choose a primary school as an example. To ensure the sanitation and safety of the population, there is no doubt that schools must be taken into consideration. In addition, children and teenagers exhibit high attack rates in epidemic outbreaks such as influenza (Frost, 1920; Monto et al., 1969) and tuberculosis (Frost, 1939). Schools also play an important role in disseminating epidemic diseases due to their close link with families, which are the basic units of the society (Jordan, 1960; Monto et al., 1969; Potter et al., 2012). Furthermore, schools can be considered as a typical population-clustered community because they carry all of the community characteristics needed for epidemic surveillance. A study of the epidemic surveillance problem in schools can improve the understanding of such problems in other communities, such as labor-intensive factories.

So far, we have seen many works related to modeling the spread of infectious diseases. Usually, the spread model uses the individuals' information as mitigation, person-to-person interaction intensity to explain the way epidemics spread. They build different deterministic or stochastic models to identify the factors which may influence the patterns or trends of spread. For example, Dye and Gay, 2003, studied how and why SARS spread from one person to another. However, the focus of this work is *not* to model SARS or any disease's spread mechanism, but on *outbreak diagnosis*, that is, to find out when and where the infection occurred. What is common is that we also believe the person-to-person interactions could potentially bring infection (location factors may do this, too). We assume all the infections happen in the common locations we identified. Hence, we collect the samples' detailed activities information and model these data to make diagnosis. This model doesn't try to create further understanding about the mechanism of spread; it instead serves as a tool to identify when and where the transmission happens, and provides information for medical investigation and control efforts.

When an alarm is triggered by an increasing number of infected subjects in a school, follow up actions should be taken to respond to early signs of an epidemic outbreak. The activities of all students and teachers (referred to as "subjects" hereafter) involve both time and location information. Therefore, such information should be recorded, and clues to possible transmission sources should be identified. The paths of subjects in a school as they move from one place to another are also informative. Information of both infected and healthy subjects is necessary to diagnose the epidemic disease. For example, an individual may be infected on Friday in school based on the incidence information, but whether the individual was infected in a classroom or on the playground, in the morning or in the

afternoon, is unknown. If the information of individuals with historically similar or different activities is given, the differences/similarities may provide clues to the transmission paths of the disease. Therefore, different from the traditional monitoring methods that only consider the number of infected cases, this work considers the daily activities of all subjects to make more focused and informative conclusions and suggestions. As such information involves both spatial and temporal details, we develop a new scheme to trace potential transmission paths and identify high-risk locations.

In the following sections, we first build a model for a community (for example, a school), namely, the expanded spatial-temporal (EST) model. As previously emphasized, multi-dimensional information should be collected for epidemic surveillance. The EST model incorporates all related information using a two-dimensional matrix, which can be handled by a wide range of existing statistical methods. Specifically, we treat each subject as one sample; the spatial and temporal dimensions of the subjects are expanded into a large one-dimension multi-way matrix. The details of the EST model are introduced in the subsequent section.

Different from the traditional studies that try to model the spread of a disease, this work is devoted to the diagnosis of infectious diseases rather than modeling their transmission. The spread model usually takes spread as "growth," individuals' location and migration as "contact distribution." Those works try to reveal how the individuals' contact information affect the trends and phenomena of the geographical spread (Mollison (1977)). However, in this work, we treat the infection status as a response variable and locations (plus time information) as explanatory variables; the diagnosis of an alarm is then equivalent to selecting variables to explain the behavior of the response variables. In our model, we assume that such transmission is caused by person-to-person interactions in certain time/location, then focus on how to use the samples' detailed activities' information to make the outbreak diagnosis. Such information would be helpful to further medical diagnosis and immunization measures.

The rest of this paper is organized as follows: Sec. 2 introduces the EST model used for diagnosis and presents the statistical framework for solving this problem; Sec. 3 investigates the performance of the proposed epidemic disease diagnosis scheme. Finally, Sec. 4 concludes this paper with suggestions for future research.

2. The Expanded Spatial–Temporal Model for Epidemic Surveillance

The extraction of diagnostic information in an infected community requires tracking the detailed activities of each subject. To characterize the behavior of a group of subjects, three-dimensional information is necessary: subject, location, and time. However, the visiting time of different subjects may not align with each other and often overlap. In this section, we first explore the features of an available dataset; then we construct a generic diagnosis model. All population-clustered communities, such as factories, have a similar information structure and data characteristics. Therefore, the methodology developed in this work can be generalized to other scenarios for the same surveillance purpose.

2.1. *Problem Description*

We introduce the problem and construct the model based on a simplified but representative infrastructure: a school with three classrooms, a library, a canteen, a study room, an office

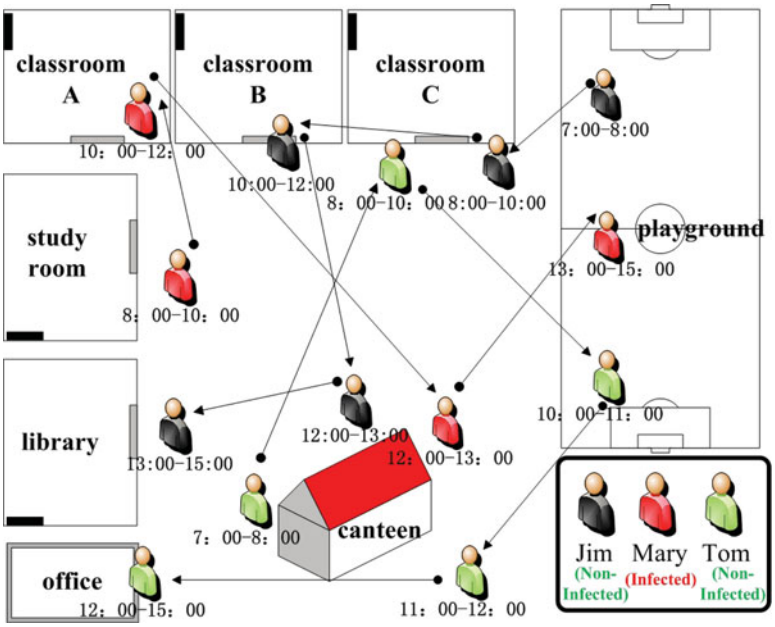


Figure 1. The information stream of daily activities in a school. (Color figure available online.)

and a playground. i.e., eight locations in all. Then we sample three subjects, Jim, Mary and Tom, and track their movement and interactions over one day, as shown in Fig. 1. Suppose that, initially, only Mary is infected. From Fig. 1, we can see the complexity of their visiting paths and interactions. Any interaction with Mary, which consists of visiting the same location at the same time, may result in an infection, depending on the infectivity of the epidemic and other factors. For simplicity, we focus on their activities in school and ignore the travel time spent between locations. If the effect of travelling on infection cannot be ignored, travel paths can be treated as locations and can be modeled in the same way.

From Fig. 1, we see that different subjects may visit one location at different time periods; the lengths of the visits may also be different. Two major challenges exist when characterizing the above scenario:

(1) The time scale is continuous whereas the locations are discrete. How can we integrate these two types of information?
(2) The visits of different subjects to one location may overlap with each other in time. How can we distinguish the subjects in the same time period?

These two challenges are unique but essential in epidemic diagnosis. Existing outbreak detection methods that use aggregated data do not need to handle such issues. In the following sections, we propose an expanded matrix to encompass a complex dataset and to allow us to find the infection sources.

2.2. *Data Manipulation*

Figure 1 demonstrates an intuitive geographic structure of the information. We now transform the data in the figure to a Gantt-type chart, as shown in Table 1. If a subject visits a

Table 1
Timetable for the daily activities of the subjects at school

Name	Mary	Tom	Jim
7:00–8:00		Canteen	Playground
8:00–9:00	Study Room	Class C	Class C
9:00–10:00			
10:00–11:00	Class A	Playground	Class B
11:00–12:00		Canteen	
12:00–13:00	Canteen	Office	Canteen
13:00–14:00	Playground		Library
14:00–15:00			

certain location during a specific time period, we annotate the location on the corresponding segment of the timetable. For the purpose of illustration, we assume that location changes occur only on the hour. For a general situation in which such changes are allowed on a shorter time interval, more segments will be generated and the scale of the problem will increase; the proposed scheme for handling such problems is still functional.

Based on the data structure defined in Table 1, we can always find time intervals during which no subjects visit two locations. In this example, the interval is one hour. This means that if each segment is taken as one hour, we can segment the continuous time into eight discrete time intervals without interfering with the location arrangement. In this way, all of the subjects share the same time segmentation. In a more general situation, we can always segment the time without having conflicts with the location arrangement; the intervals do not need to have the same length. This allows the continuous time to be transformed into discrete time intervals. If the segment length decreases, the number of segments would increase, and the size of the timetable (a two-dimensional matrix) would grow. As aforementioned, the proposed surveillance scheme would still functional.

Next, we combine the segmented time intervals and locations. The spatial-temporal plot shown in Fig. 2 is created to describe the data from a two-dimensional view. In Fig. 2, the visiting path of each subject is a piecewise function. We can take one location at a time interval as a *spatial-temporal point*. For example, the playground from 7:00 to 8:00 can be considered as a spatial-temporal point. Each subject can visit only one location during one time interval. Hence, the visiting status of a subject at each spatial-temporal point is described by a binary variable: visiting or not. By defining each spatial-temporal point (i.e., the combination of location and time) as one variable, we can further transform the information shown in Fig. 2 to Table 2.

In Table 2, each column title represents a time-location combination. For example, x_1 represents (7:00–8:00, classroom A), x_2 represents (7:00–8:00, classroom B), and x_9 represents (8:00–9:00, classroom A). For each combination, the subject may or may not be present. An indicator value of 1 and 0 is then used to denote whether the subject was present or not, respectively. Using these notations, we transform the three-dimensional data (time,

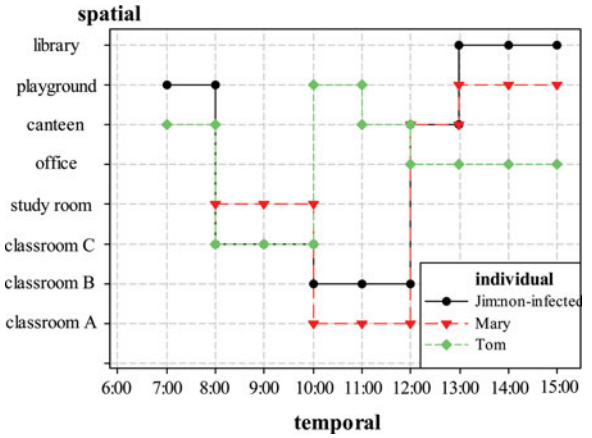


Figure 2. The spatial-temporal data on a coordinate graph. (Color figure available online.)

location and subject) into a two-dimensional matrix (time × location as a single dimension and subject as another dimension). Because the new matrix is formed by expanding the spatial and temporal information of the subjects, we name it an expanded spatial-temporal model.

To identify infection sources, the (time, location) combinations could be viewed as predictors or explanatory variables; the response variable y indicates the infection status of a subject. Each row, therefore, contains the complete path of a subject within a definite time range.

Thus, the complete spatial and temporal information has been expanded into a two-way matrix. When infection cases are found and we are interested in identifying the sources of infection, it is equivalent to identify which variables (or a single variable) have significant impact on the value of the response variable y. Since the time and location data have been integrated with the explanatory variables, when such variables are located, we can easily attribute the outbreak to specific time and location combinations. For example, if all subjects with $x_1 = 1$ are infected, we may strongly suspect that epidemic infection occurred in Classroom A during 7:00–8:00 am. Therefore, the original diagnosis problem is translated into a statistical problem of identifying variables among predictors that have significant effects on the infection status.

Table 2
The extracted X and Y variables under each record

Subject	x_1	x_2	x_3	x_4	x_5	... ...	x_{63}	x_{64}	y
1	0	0	0	0	0	... ...	1	1	0
2	0	0	0	1	1	... ...	0	0	1
3	0	0	0	0	0	... ...	0	0	0

2.3. *The EST Modeling Framework*

Let $\mathbf{x} = (x_1, \ldots x_i, \ldots, x_K)^T$ be a vector of all predictors, and

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iK} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nj} & \cdots & x_{NK} \end{pmatrix}$$

be a matrix containing the trajectories of all subjects being studied, and $\mathbf{y} = (y_1, \ldots y_i, \ldots, y_N)^T$ be a vector containing the infection status of all subjects. If the ith subject is infected, $y_i = 1$; otherwise, $y_i = 0$. The complete information for the diagnosis is therefore encapsulated in $\mathbf{X}$ and $\mathbf{y}$.

The total number of (time, location) combinations K is determined by the size of the campus and the fineness of the time domain being considered. For a target community with L locations, if the time horizon is divided into S segments, the total number of explanatory variables K would be $K = L^*S$. Therefore, the formulation framework proposed in this work is applicable to problems of any size. Naturally, if K increases, the number of subjects needed for an accurate identification would increase accordingly.

In order to find the variables in $\mathbf{x}$ that have a significant effect on y, we need to build a model between these two sets of variables. Because y is a binary variable that can be 1 or 0, a logistic regression model would be convenient for building the relations among all variables. Using the above notations, the probability that one subject would be affected based on a given visiting path (spatial-temporal points and visiting history) can be expressed as follows:

$$P(\mathbf{y} = 1|\mathbf{X}) = \pi(\mathbf{X}) = \frac{1}{1 + e^{-g(\mathbf{X})}} \tag{1}$$

where

$$g(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{2}$$

The coefficient vector, $\boldsymbol{\beta}$, reflects the significance of each predictor on $\mathbf{y}$. The solution of Eq. (2) will be introduced in the following section.

We now briefly summarize our proposed modeling strategy for epidemic diagnosis. For population-clustered communities, such as the school in our example, the daily spatial and temporal information of subjects could be collected. At the same time, the infection status of each subject is assumed to be known. The spatial and temporal data are unfolded and sliced into small segments. The spatial-temporal points are treated as explanatory variables $\mathbf{x}$. Using the infection information represented by $\mathbf{y}$, an alarm can be back-traced by searching for a subset of $\mathbf{x}$ variables that have a significant impact on $\mathbf{y}$. This modeling process transforms the epidemic diagnosis problem into a regression model, and in the next section, we will introduce variable-selection methods to determine the $\mathbf{x}$ variables that have a significant impact on $\mathbf{y}$.

It is emphasized that the availability of detailed information of individuals is critical to the EST model. In practice, collection of such information from existing IT systems could be considered, such as the registration, accommodation, dinning, and curriculum systems

in a school. As aforementioned, for new and high-risk infectious diseases, the collection of such information may be necessary even additional cost need to be paid.

In addition, as the focus of this work is to provide diagnostic information regarding when and where is the location/time that people get infected. After identifying such information, further analysis with the help of medical domain knowledge is needed. The incubation period and the medical recognition and confirmation of a case do not affect our ***statistical diagnosis*** model. For example, suppose the incubation period of a particular disease is 10 days. When trying to identify the transmission place for patients confirmed on April 20, we should analyze the data back to April 10 and identify the common time/location they have visited. Instead, if the incubation period is 5 days, we should look at the trajectories of the patients on April 15. Due to observational noise (e.g., delayed diagnosis), it should be beneficial to cover more data in the analysis.

2.4. *Diagnosis via Variable Selection*

So far, we have translated the diagnosis problem into an EST model by segmenting the time axis and combining the spatial and temporal information. Next, we need to identify the variables in $\mathbf{x}$ that have a significant impact on $\mathbf{y}$. Automatic variable selection is an important topic in data analysis. Traditional methods, such as ordinary least squares (OLS) estimation, subset selection and stepwise regression, are applicable if we want to find the significant variables in Eq. (2). However, these methods are less favored due to stability or efficiency issues.

Tibshirani (1996) proposed a least absolute shrinkage and selection operator (Lasso) method for variable selection. The Lasso method attempts to estimate the value of β by solving the following minimization problem:

$$\min \sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 s.t. \sum_j |\beta_j| \leq t$$

where t is a regulation parameter. A positive t has the effect of forcing certain small coefficients in vector β to exactly zero, consequently, removing insignificant variables from Eq. (2) and increasing the interpretability of the model.

However, the traditional Lasso algorithm assumes that the response variable is continuous. To incorporate logistic regression with variable selection, algorithms such as Bayesian binary regression (Genkin et al. (2007)), the l1logreg algorithm (Koh et al. (2007)) and glmnet algorithm (Friedman et al. (2010)) have been developed. The glmnet method is a logistic regression with sparse features that is faster than the other two methods. Therefore, in this work, the glmnet algorithms employed in these work to select explanatory variables from Eq. (2) and provide diagnostic information for surveillance.

The glmnet algorithm maximizes the following penalized log likelihood function

$$\max_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{N} \sum_{i=1}^{N} \{ I(y_i = 0) \log p(x_i) + I(y_i = 1) \log(1 - p(x_i)) \} - \lambda P_\alpha(\beta) \right] \quad (3)$$

where

$$P_\alpha(\beta) = \sum_{j=1}^{K} \left[\frac{1}{2} (1 - \alpha) \beta_j^2 + \alpha |\beta_j| \right]$$

is the elastic-net penalty, which sets the coefficients of the insignificant explanatory variables to zero. The pseudo code of the glmnet algorithm for variable selection is provided in the appendix. More details of this algorithm can be found in the original work of Friedman et al. (2010).

The solution to Eq. (3) can reveal the variables that are significant to the infection status. Variables with nonzero coefficients may potentially be responsible for epidemic infections. Because each variable is defined by a combination of location and time, such information provides further insights for diagnosing the root cause.

3. Performance Study

To validate the EST model and evaluate its performance, we conduct studies based on data from a simulated school in this section. Here, we still use *point* to denote a spatial-temporal combination, which is also a variable in the EST model. An infection point is a point that has contributed to the spread of an epidemic. The following two indices are used in the studies for performance evaluation:

(1) False discovery rate (FDR): the proportion of points that are wrongly identified as infection sources. If no sources of infection are identified, FDR is zero.
(2) Omission rate (OR): the proportion of points that are in fact infection sources but that are missed by the selection algorithm. If all points are identified as sources, OR is zero.

The following parameters, which may vary with the school size and the properties of different epidemics, can influence the performance indices:

(1) Infectious rate: this rate represents the infectivity of a specific epidemic disease,
(2) λ: this is the statistical tuning parameter shown in Eq. (3); it affects the weight of the penalty in the objective function and influences the number of variables that can be selected.

In the following sections, these parameters are studied under different settings; the corresponding performance results are reported for comparison.

3.1. *Simulation Setup*

The simulation is based on the model described in Sec. 2, which is a school with a certain number of subjects and locations. The following settings are used in the simulation:

(1) The number of tracked subjects, which represents the number of rows in $\mathbf{X}$. In this simulation, three levels for this number are studied: a small class level with 30 subjects, a large class level with 50 subjects and a grade level with 300 subjects.
(2) The number of points (spatial-temporal combinations), which is the number of variables in the model. For the case with 30 subjects, we assume there are 10 points; for the case with 50 subjects, we assume there are 20 points; for the case with 300 subjects, there are 100 points.
(3) The infection points, which are defined as points that are infectious. If a non-infected subject visits an infectious point, the subject has a chance of being infected. These infection points are the true root causes that we want to identify in diagnosis. It should be noted that, an infectious disease is commonly transmitted from one subject to other subjects. Therefore, it is the subject, rather than the point,

that is infectious. The infectious points should be dynamic, following the travelling path of the infected subjects. However, for a given visiting history, we can consider any spatial-temporal combination that contains infected subjects at infection points as points at which the disease is potentially spread. This treatment only simplifies the generation process of simulated data and does not change the data structure or the implementation of the proposed method. In the following study, we assume that there are 3 infection points for the case with 10 points, 5 infection points for the case with 20 points and 12 infection points for the case with 100 points.

(4) The visiting rate to the infectious and non-infectious points. The travelling path of each subject is assumed to be random, independent, and non-repeatable (since a subject could not re-visit a point that contains the same temporal information). The visiting rate defines the probability that a location is visited by the subjects. For illustration, we set the visiting rate to the non-infectious points to be 0.3 and the rate to the infectious points to be 0.4. Then, each non-infectious point will be visited by the subjects at a probability of 30%, and each infectious point will be visited by the subjects at a probability of 40%.

(5) The infectious rate: after visiting an infectious point, a subject is infected at the infectious rate. This parameter is set to 0.3 for low-level infectious diseases, 0.5 for middle-level infectious diseases and 0.7 for high-level infectious diseases.

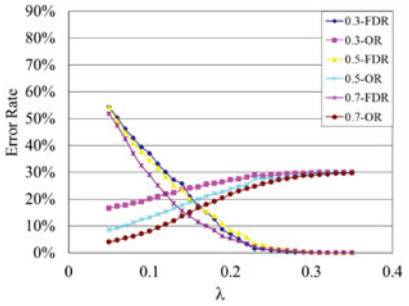
In total, nine scenarios are used for the simulation, as shown in Table 3. Each performance index is calculated based on the results from 100 replicates.

3.2. *Performance Comparison*

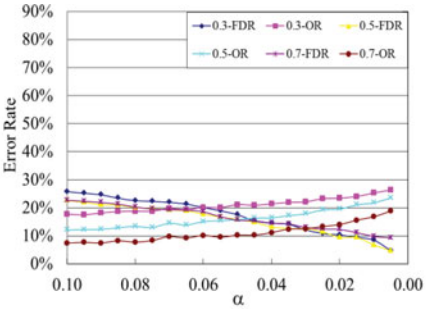
In the current practice of epidemic diagnosis, a natural and simply way to identify root causes is to perform a statistical test for two proportions. For example, to verify whether x_1, which corresponds to a specific location and time, is a point that influences the infection status, one may classify all subjects into two groups: those with $x_1 = 1$, i.e., those who have this point in their travelling path, and those with $x_1 = 0$, i.e., those who do not have this point in their visiting path. Then, one may use a test for two proportions to determine whether the numbers of infected subjects are significantly different. If they are, this point is considered the reason for the difference and, is treated as an infection point.

Table 3
Settings for simulation

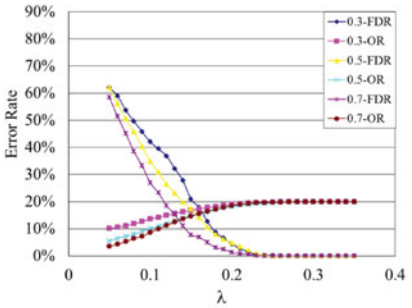
Scenario	# of Subjects	# of Points	# of Infection points	Infectious rate
1	30	10	3	0.3
2	30	10	3	0.5
3	30	10	3	0.7
4	50	20	4	0.3
5	50	20	4	0.5
6	50	20	4	0.7
7	300	100	12	0.3
8	300	100	12	0.5
9	300	100	12	0.7



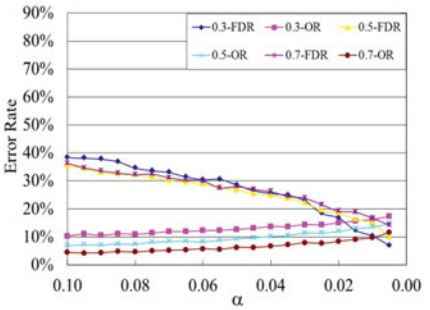
(a) Lasso variable selection. Subjects = 30, points = 10, infection points = 3. Infectious rate = 0.3, 0.5 and 0.7.



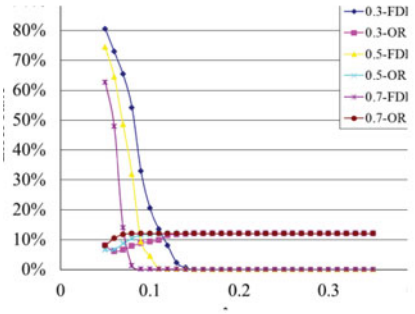
(b) 2-proportion test. Subjects = 30, points = 10, infection points = 3. Infectious rate = 0.3, 0.5 and 0.7.



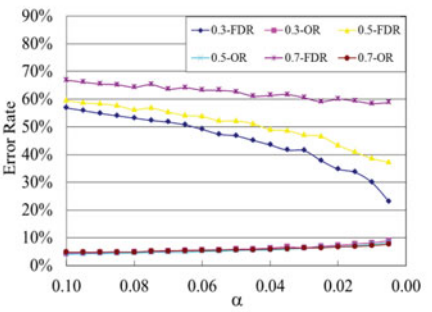
(c) Lasso variable selection. Subjects = 50, points = 20, infection points = 4. Infectious rate = 0.3, 0.5 and 0.7.



(d) 2-proportion test. Subjects = 50, points = 20, infection points = 4. Infectious rate = 0.3, 0.5 and 0.7.



(e) Lasso variable selection. Subjects = 300, points = 100, infection points = 12. Infectious rate = 0.3, 0.5 and 0.7.



(f) 2-proportion test. Subjects = 300, points = 100, infection points = 12. Infectious rate = 0.3, 0.5 and 0.7.

Figure 3. Performance comparison. (Color figure available online.)

In the following, the 2-proportion test procedure is used as an alternative to our proposed method for diagnosis, and the results from the two methods are compared by calculating the same criteria, i.e., the false discovery rate and the omission rate. The results are shown in Fig. 3. The numbers of subjects, total points, infection points, and the method used, are shown beneath each plot.

It should be noted that for the methods using Lasso variable selection based on the proposed EST model, when the tuning parameter λ in Eq. (3) increases, more zeros will be seen in the coefficient vector after variable selection. Therefore, OR will increase accordingly, as FDR decreases. For the 2-proportion test, when type I error rate α increases, there is a trend that more points will be judged as infectious. Consequently, OR will decrease as FDR increases. For comparison, we arrange the horizontal axis of the 2-proportion test to decrease so that the FDR and OR values on both plots would show the same increasing or decreasing trend.

It is learned from Fig. 3 that:

(1) For both of the methods studied, FDR decreases and OR increases when λ increases or α decreases. The Lasso-based method shows a sharp decreasing trend in the FDR curve when λ increases from zero, while a much slower increasing trend is observed in the OR curve. When λ exceeds a certain value (approximately 0.2), both curves become flat. This suggests that an optimal λ can be chosen to achieve a balance between FDR and OR. Similar patterns are observed for the 2-proportion test methods, except that the decrease in FDR is much slower.
(2) As shown in plots (a) and (b), a lower FDR and a higher OR are observed when the Lasso variable-selection algorithm is used, as compared to the 2-proportion test. In plot (b), the cases with α equal to approximately 0.03 or 0.05 achieves an overall good performance.
(3) Plots (c) and (d) shows that the FDR of the 2-proportion test becomes rather high for a problem with a moderate scale. When the scale of the problem becomes larger (as shown in plots (e) and (f)), the FDR of the 2-proportion test becomes unacceptably high, even for a very small α. On the other hand, the FDR of the Lasso variable-selection method is almost zero at $\lambda \geq 0.15$; the corresponding OR is stabilized at approximately 11%.
(4) It is also found that when the infectious rate increases, both the FDR and OR of the Lasso variable-selection method tend to decrease; however, the FDR of the 2-proportion test becomes larger as the infectious rate increases from 0.3 to 0.7.

We therefore conclude that the 2-proportion test method is competitive for small-scale scenarios. However, for moderate- or large-scale scenarios with hundreds of subjects and points (spatial-temporal combinations), the 2-proportion test method becomes less efficient. Instead, the Lasso variable-selection method is more advantageous and stable for large-scale scenarios. The Lasso-based variable-selection scheme can identify spatial-temporal information regarding the sources of infection, thereby providing guidelines for effective immunity efforts.

4. Conclusions

Epidemic surveillance in a community involves both outbreak detection and source diagnosis. Even though identifying the disease transmission path in a timely manner is essential for effective immunity efforts, most existing works have focused on detecting outbreaks based on aggregated data. When the complete travelling paths of subjects (including both location and time information) are known, making full use of such data for diagnosis becomes challenging.

In this paper, we propose an expanded spatial-temporal (EST) model to characterize the activities of activities of the subjects. Three-dimensional information (subject, location and time) is expanded into a two-dimensional space by dividing the time horizon into

segments and by multiplying the segments with the locations. Based on the EST model, we further propose to use a variable-selection algorithm to identify potential location/time combinations as transmission sources and to satisfy the goal of diagnosis. Numerical simulations show that the proposed scheme is effective in locating sources of infection.

The EST model can be utilized to illustrate any general communities with additional subjects and different types of activities. Therefore, although motivated by a school example, the proposed diagnosis scheme can be extended to more general scenarios to help control epidemic transmission in practice by using more effective measures.

Current research on outbreak detection mainly relies on the counted numbers of subjects at certain aggregated levels. Because the EST model contains more detailed information on the subjects as compared to traditional methods, detection methods based on the EST model are a promising technique that merits future research efforts.

Appendix

Pseudo Code for Lasso Variable Selection

The glmnet package developed by Friedman et al. (2010) for variable selection uses a coordinate descent algorithm to find the minimization of the above objective function; the whole solution path for $\lambda_{\min} \le \lambda \le \lambda_{\max}$ is calculated ($\lambda_{\min}$ is defined below). Let x_i be the ith observation, y_i be the ith response, N be the total number of observations, β_j be the jth component of β, and $(\tilde{\beta}_0, \tilde{\beta})$ be the estimate of (β_0, β). The algorithm works as follows (for more details of this algorithm, please see Friedman et al. (2010)):

1) First, obtained the smallest $\lambda_{\max}$ such that $\tilde{\beta} = 0$, that is, find the smallest weight for the penalty such that all estimated elements are penalized to zero;
2) Determine $\lambda_{\min} = \varepsilon\lambda_{\max}$ as the lower bound for the tuning parameter λ. The parameter ε usually takes the value of 0.001.
3) For $k = 0, 1, 2, \ldots, K$ (K typically equals 100)
   a) Decrease λ: $\lambda = \varepsilon^{\frac{k}{K}} \cdot \lambda_{\max}$
   b) For all $i = 1, \ldots, N$, calculate $\tilde{p}(x_i) = (1 + e^{-(\tilde{\beta}_0 + x_i^T \tilde{\beta})})^{-1}$, which is the probability that the response variable takes the value 1;
   c) For all $i = 1, \ldots, N$, calculate $w_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i))$
   d) For all $i = 1, \ldots, N$, calculate $z_i = \tilde{\beta}_0 + x_i^T \tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))}$
   e) Update $\ell_Q(\beta_0, \beta) = -\frac{1}{2N} \sum_{i=1}^{N} w_i (z_i - \beta_0 - x_i^T \beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2$
   f) Solve $\min_{(\beta_0, \beta) \in \mathbb{R}^{P+1}} \{-\ell_Q(\beta_0, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}$ by coordinate descent to get $(\tilde{\beta}_0, \tilde{\beta})$, thus obtain the solution corresponds to the specific λ assigned in step a).

Acknowledgments

The authors thank the editor and the anonymous referees for their helpful comments, which have helped improve this work greatly. This work was supported by the Natural Science Foundation of Beijing under grant No. 9092005.

References

Allard, R. (1998). Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization* 76(4):327–333.

Andersson, E. (2009). Effect of dependency in systems for multivariate surveillance. *Communications in Statistics-Simulation and Computation* 38(3):454–472.

Bailey, N. T. J. (1986). Macro-modelling and prediction of epidemic spread at community level. *Mathematical Modelling* 7(5–8):689–717.

Carey, R. G. (2003). *Improving Healthcare with Control Charts: Basic and Advanced SPC Methods and Case Studies*. American Society for Quality.

Cash, R. A., Narasimhan, V. (2000). Impediments to global surveillance of infectious diseases: Consequences of open reporting in a global economy. *Bulletin of the World Health Organization* 78(11):1358–1367.

Cowden, J., Lynch, D., Joseph, C., O'Mahony, M., Mawer, S., Rowe, B., Bartlett, C. (1989). Case-control study of infections with Salmonella enteritidis phage type 4 in England. *British Medical Journal* 299(6702):771.

Dye, C., Gay, N. (2003). Modeling the SARS epidemic. *Science* 300(5627):1884–1885.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Frisen, M., Wessman, P. (1999). Evaluations of likelihood ratio methods for surveillance. Differences and robustness. *Communications in Statistics-Simulation and Computation* 28(3):597–622.

Frost, W. (1920). Statistics of influenza morbidity: with special reference to certain factors in case incidence and case fatality. *Public Health Reports* (1896–1970): 584–597.

Frost, W. H. (1939). The age selection of mortality from tuberculosis in successive decades. *American Journal of Epidemiology* 30(3):4–9.

Genkin, A., Lewis, D. D., Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* 49(3):291–304.

Hanslik, T., Boelle, P. Y., Flahault, A. (2001). The control chart: An epidemiological tool for public health monitoring. *Public Health* 115(4):277–281.

Horby, P., O'Brien, S., Adak, G., Graham, C., Hawker, J., Hunter, P., Lane, C., Lawson, A., Mitchell, R., Reacher, M. (2003). A national outbreak of multi-resistant Salmonella enterica serovar Typhimurium definitive phage type(DT) 104 associated with consumption of lettuce. *Epidemiology and Infection* 130(2):169–178.

Hu, S., Li, J., Deng, Z., Li, Y., Zhang, H., Long, Z., Liu, Y., Guo, S. (2004). Investigation report of six SARS cases in hunan province. *Chinese Journal of Disease Control and Prevention (in Chinese)* 8(3):275–276.

Hutwagner, L., Thompson, W., Seeman, G. M., Treadwell, T. (2003). The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 80(Supplement 1):i89–i96.

Jordan Jr, W. (1960). Stability characteristics of Goe virus. *Proceedings of the Society for Experimental Biology and Medicine* 103(3):506–9.

Koh, K., Kim, S. J., Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research* 8(8):1519–1555.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164(1):61–72.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society. Series B (Methodological)* 283–326.

Monto, A. S., Davenport, F., Napier, J., Francis Jr., T. (1969). Effect of vaccination of a school-age population upon the course of an A2/Hong Kong influenza epidemic. *Bulletin of the World health Organization* 41(3-4-5):537–542.

Pfeiffer, D. U., Minh, P. Q., Martin, V., Epprecht, M., Otte, M. J. (2007). An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal* 174(2):302–309.

Potter, G. E., Handcock, M. S., Longini, I. M., and Halloran, M. E. (2012). Estimating within-school contact networks to understand influenza transmission. *The Annals of Applied Statistics*, 6(1), 1–26.

Rolka, H., Burkom, H., Cooper, G. F., Kulldorff, M., Madigan, D., Wong, W. K. (2007). Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs. *Statistics in Medicine* 26(8):1834–1856.

Shmueli, G., Burkom, H. (2010). Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics* 52(1):39–51.

Sonesson, C., Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 166(1):5–21.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Tsui, K. L., Chiu, W., Gierlich, P., Goldsman, D., Liu, X., Maschek, T. (2008). A review of healthcare, public health, and syndromic surveillance. *Quality Engineering* 20(4):435–450.

WHO. (2010). Pandemic (H1N1) 2009 - situation report update 100. Available at http://www.who.int/csr/don/2010_05_14/en/index.html

Woodall, W. H., Brooke Marshall, J., Joner Jr, M. D., Fraker, S. E., Abdel-Salam, A. S. G. (2008). On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(1):223–237.

Yip, P. S. F., Lam, K. F., Xu, Y., Chau, P. H., Xu, J., Chang, W. H., Peng, Y. C., Liu, Z. J., Xie, X. Q., Lau, H. Y. (2008). Reconstruction of the infection curve for SARS epidemic in Beijing, China using a back-projection method. *Communications in Statistics-Simulation and Computation* 37(2):425–433.